Introduction to Phylogenetics

Lucie Abeler-Dörner

With Next Generation Sequencing (NGS) becoming more affordable every year, it has become feasible to sequence large numbers of samples in a quality that was out of reach even a decade ago. How can sequences complement classical epidemiology? What insights does it give us that classical epidemiology cannot? What are its limitations? What are the ethical considerations? These are some of the questions that this factsheet will aim to explore.

Classical epidemiology tells the story of an epidemic from the point of view of the host whereas phylogenetics tells the story from the point of view of the virus. In many instances, the results can complement and verify findings obtained by classical epidemiological methods, sometimes the same findings can be obtained a lot easier and with less error, and occasionally, phylogenetics will provide insights that could not have been obtained otherwise.

Why has NGS changed the game?

NGS Illumina sequencing produces many reads of the same part of the sequence. This is referred to as the sequencing depth. If the sequencing for a particular sequence is 5 (or 5x), then each part of the sequence has been sequenced at least five times. With modern sequence technology, a sequencing depth of 30 is not unusual. With traditional Sanger sequencing, only the most common base for each position could be recoded with certainty. Such a sequence is called a consensus sequence.

With NGS, all variants and their frequencies can be recorded. These are called minority variants. It is important to note that a minority variant only refers to a change at a given nucleotide position, not to a whole variant genome. The reason for this is that during the sequencing process, the whole genome is chopped up into manageable chunks, often of a median length of around 250 base pairs (bp). The sequencing method we use in Oxford has been optimised to generate a larger proportion of long fragments of more than 350 bp, as this improves the analysis of transmission networks (see below). We do therefore not know which of the 250 or 350 bp fragments should be aligned to create the original genome.

This limitation is overcome by new protocols and new sequencing technology like Oxford Nanopore sequencing, but at the moment these systems are not high throughput and more error-prone than standard NGS sequencing. However, the field is evolving fast and we will likely see more sequences generated by these methods in the future. The big advantage of the Oxford Nanopore technology is that it does not have a restriction on the length of the fragments that it can read. So theoretically, it can read a whole HIV genome in one go. This single read of one individual genome is called a haplotype. Practically, it is however rather tricky to produce such long cDNAs in sufficient quantities.

What is phylogenetics?

Phylogenetics comprises a set of techniques to build and interpret trees that are based on the similarity of different genetic information. A fictional example of a phylogenetic tree is depicted in figure 1. This figure was shamelessly copied from Andrew Rambaut's highly recommended tutorial on how to read a phylogenetics tree: http://artic.network/how-to-read-a-tree.html.

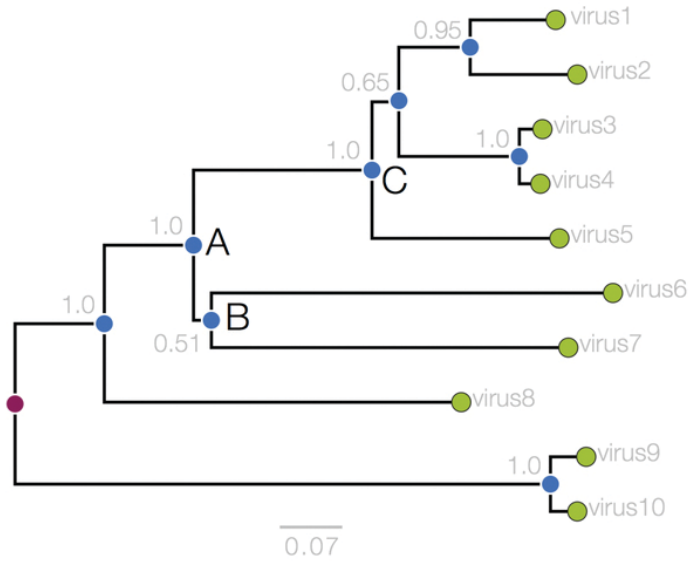

Figure 1: A typical rooted phylogenetic tree

Each genetic entity – in this case a viral sample from a single organism – is represented by a green tip at the right of the tree. The horizontal distance denotes similarity, mostly expressed as nucleotide substitutions or number of substitutions per 100 bases. Note that at this stage in the process, the horizontal dimension of the tree is not linked to time. The vertical lines on the tree do not carry any information, they are just there to make the tree more readable. The blue circles or internal nodes are common ancestors of sequences in the dataset. They are inferred and not present in the dataset themselves. The red circle is the root of the tree. There are different techniques not determine the root or not determine the root which will result in different tree layouts. See Andrew's tutorial for details. The genetic distance between two tips of the tree is the length of all horizontal lines that you have to follow to get from one tip to the other. Most phylogenetics analyses focus on the right-hand side of the tree and are mainly concerned with how the tips fall into different clusters. Phylodynamic analysis often focus more on the left-hand side of the tree.

Phylogenetic trees can not only be used to compare the similarity between different samples, they can also be used to compare differences between sequences in the same sample. Since HIV is a fast-evolving virus, variants of the original virus can be detected in the host already a few months after infection. Adding within-sample variation to the analysis will result in a tree like the one in figure 2. The sequences from one sample (denotes in the

same colour) are in most cases more closely related to each other than the sequences from different samples (denotes by different colours), but still show a large amount of measurable diversity. Information about this diversity is extremely helpful when constructing transmission networks as viral sequences of one patient falling completely within the diversity of sequences of another patient (F and E in figure 2) strongly suggests that virus from person E was directly or indirectly transmitted to person F.

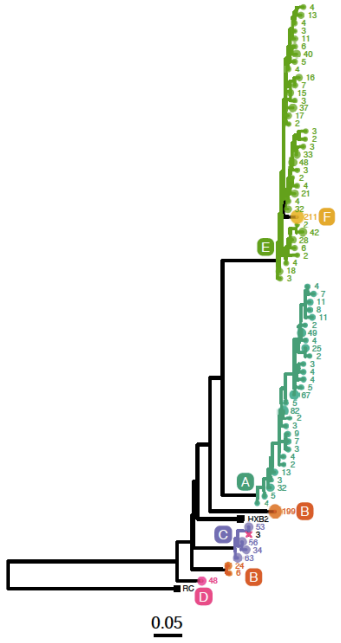

Figure 2: A phylogenetic tree depicting variation within the samples (within one colour) as well as similarity between samples (A-F, different colours, courtesy of Chris Wymant

One thing that is really important to understand about representations of trees like this is that the output of the analysis was not a single tree, but thousands of trees that were then collapsed into a single summary tree. A number near the internal nodes of the tree often indicates in which fractions of generated trees this node was present. A tree is therefore not a depiction of the truth, but of a set of probabilities. When comparing longer sequences, e.g. the whole HIV genome, it is therefore good practice to construct multiple trees from different parts of the genome to make and make sure the resulting trees are consistent. This will also detect recombination events, as different genes in a clade A-D recombinant will cluster differently with other sequences when looking at different parts of the genome.

Trees can be modified or built with additional information. Genetic similarity could be based on protein rather than nucleotide sequence, for example, and additional data, e.g. infection dates can be used to restrict the number of possible trees that are allowed. This leads us into phylodynamics.

What is phylodynamics?

Whereas phylogenetics is solely concerned with similarity (or lack thereof), phylodynamics aims to explain how these patterns arose. This requires two things: Introducing an element

of time and introducing a model that can explain patterns that have occurred in the past. The power of this approach is that using a mathematical model makes it also possible to predict developments in the future.

Time: Linking genetic similarity to evolutionary processes requires the introduction of a so-called “molecular clock” as a means to convert the horizontal lines from changes in similarity to time passed. This molecular clock is the rate at which the pathogen in question accumulates genetic changes. It depends among other things on the time the pathogen requires to complete an infection cycle and the genetic set-up of the pathogen (e.g. the error-rate of the polymerase) and can be estimated using different methods. The molecular clock can be different in different host species (of relevance for influenza for example) and is also likely to be different between individual viruses of the same strain and clade, especially when subjected to different evolutionary pressures. The assumed molecular clock applied to a tree is therefore always an estimate of the mean rate and its variance will limit the precision with which the tree can be timed.

Mathematical model: As a second element required for phylodynamics analyses is a mathematical model that aims to explain the observed patterns and is therefore able to model the continuation of these patterns into the future. These models can range from very simple, e.g. just taking into account date and location of sampling, to more complicated models which can incorporate dozens of metadata fields.

Transmission networks

If the sampling in a given population or a given outbreak is sufficiently dense (upwards of 15% of people infected), it is possible to construct transmission networks, e.g. to calculate the probability that one patient transmitted virus directly or indirectly to another patient. These relationships and probabilities are usually depicted in a network like the one in figure 3.

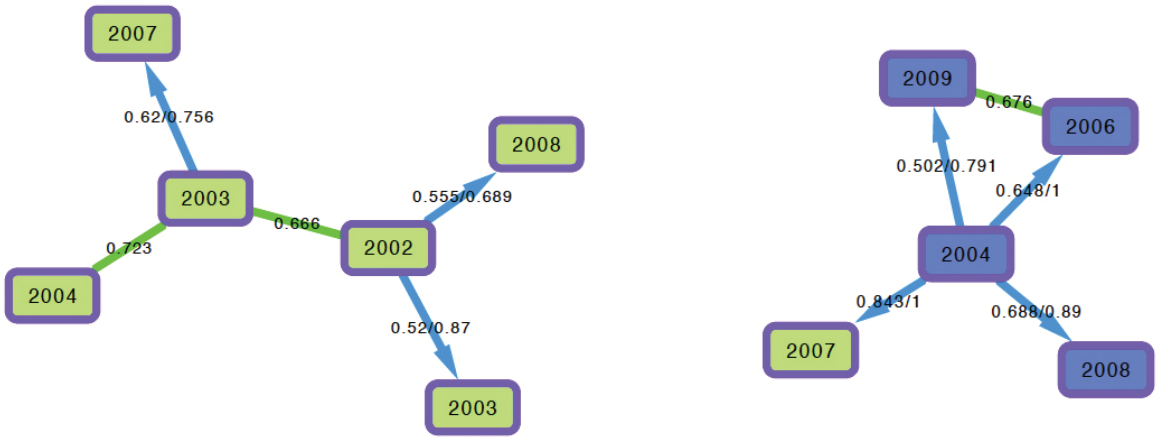

Figure 3: Typical transmission network. Each box is a patient, numbers in boxes are infection dates, number on lines are probability that these patients are linked / probability that the directionality indicated by the arrow is correct, courtesy of Matthew Hall

What questions can phylogenetics and phylodynamics answer?

Drug resistance mutations: Known drug resistance mutations can be identified in the polymerase and integrase genes and, if sequences are linked back fast enough, can inform the choice of therapy. This is the only direct benefit for patients.

Analysis of viral recombinants: The methods we use can correctly assemble and align recombinant virus. It is therefore possible to study the impact of recombination on other variables, e.g. those linked to clinical outcome.

Tracking viral evolution: Sequence data can be used to model how HIV spread, whether in the last few years in a given area or over the last 100 years world-wide. Phylodynamic analyses revealed for example that HIV was introduced times to the US on many different occasions (and that the so-called “patient zero” only played a marginal role). It is also of relevance now as public health measures can vary dependent on whether most infections occur from virus circulating within a population or by continuous introductions from outside.

Better measures of incidence: In study areas affected by high migration, it is often difficult to assess if migrants were infected before they moved into the area or if they were infected within the study area. Phylogenetics improves the measurement if incidence in this case, as viruses distinct from those found usually in the study population suggest that infection of a migrant occurred before their arrival in the area.

Early detection of outbreaks: Within a generalised epidemic it is often difficult to assess if a local increase is s genuine change of transmission dynamics in this area or group or just fluctuation. Analysing the sequences can help to assess if rapid transmission is occurring in a group of people within the generalised epidemic.

Information on who infected whom: Transmission networks are a powerful tool to assess how the virus is spreading in a population and is less error-prone, less time-consuming and less costly than contact tracing. There are however ethical concerns about generating this highly sensitive information and anonymity of the participants needs to be ensured at all times.

Identification of risk groups: Groups with high prevalence can be the source of new infections, but also the sink for infections from other parts of the network. Transmission analysis can identify groups who are most at risk of being infected and groups that are most of risk of infecting others.

Identification of super-spreaders: Transmission networks can reveal individuals who cause a disproportionally high number of new infections. This is relevant for targeted prevention as well as for adjusting mathematical models of incidence which are often susceptible to changes in transmission homogeneity.

Identification of infection cycles: Taking all the above information obtained into account, we can ask if cycles of patterns of infection can be observed that explain how virus is passed on from one generation to the next.

Identification key groups for effective prevention: The ultimate goal for the Gates Foundation is to identify groups that should be targeted with preventive measures (e.g. injectable PreP to have a maximum impact on the course of the epidemic. This can be achieved by assessing the contribution of different groups to the epidemic and targeting those groups that have been shown to have a disproportionate role in viral transmission (and are on board with doing something about it).

Modelling of different prevention packages: Finally, mathematical models can be developed that aim to explain the course of the epidemic as revealed by the sequence data. Prevention packages could then be simulated by reducing the incidence in those groups that are targeted by the envisaged prevention. It all sounds very logical, but in fact is far from straight forward, so will still require a lot of work.

Limitations and dependencies

Error rate: Current NGS technology usually has an error rate of approximately 0.5%. This is not an issue when working with consensus sequences, but can lead to errors when looking at minority variants and haplotypes

Amplicon length: For certain types of phylogenetic analysis, it is particularly useful to have long reads, or even complete haplotypes. Some of the sequences generated during PANGEA 1 have short reads that are challenging for some of the analyses.

Study design: Most of the time, sequencing is an add-on to an existing study and the study-design might therefore not be ideal. For example, if the study aimed to sample one member of each house-hold, many partners will be missing in the transmission network. If the study sampled broadly, e.g. like PHIA, this is ideal for many phylodynamics analysis, but will not be suitable for transmission networks as the data is too scarce. Conversely, when certain risk groups were studied, the dataset is ideal for transmission network analysis but not for answering phylodynamics questions. When comparing datasets that were sampled with different methods, it is important to adjust for sampling bias to reach sound conclusions.

Dependency on metadata: Like all clinical data, sequence data is only useful with associated metadata. For some phylodynamics analyses as little as months and place of sampling is enough, for other analyses, e.g. migration or transmission studies, the value of the dataset grows exponentially with the amount of associated metadata.

Dependency on control data: We can obviously only obtain HIV sequences from infected individuals, but for many analyses it is important to know which proportion of the HIV positive population was included and whether these samples were preselected as the trees and networks will be biased if they are not adjusted for unequal sampling. For other analyses, especially the ones that the Gates Foundation would like us to do, it is also

important to have metadata on a representative sample of the HIV negative population in the study area as accurate data for public health can only be generated in the context of the population as a whole.

Ethical considerations

NGS data poses several challenges that were not as acute when traditional sequencing would only yield a consensus sequence. These lie particular in the area of patient identifiability and the reconstruction of transmission networks.

Patient identifiability: Since NGS can identify minority variants, the resulting set of sequences as processed by *phyloscanner* is very distinct and differs from patient to patient, almost like a viral fingerprint. Different samples from the same patient submitted to public repositories by different studies could therefore be linked and identified to belong to the same person. Equally, if a sample was taken from a person, information from this sample could be used to identify any previous samples submitted to public databases. The Executive Committee has therefore taken the decision to withdraw raw sequence files from public repositories and only submit consensus sequences which are too generic to be linked back to an individual.

Transmission networks: Transmission networks aim to establish how different patients are linked via their viral sequences. Currently, each transmission even is associated with a probability and the data is consistent with a scenario in which virus was not transmitted directly from one individual in the network to another, but via unsampled intermediaries. However, the techniques are being refined and soon it will likely be possible to state with a very high certainty that one person infected another person directly. Combining this with evidence that the two people in question had direct sexual contact in a period that is consistent with the estimated window of transmission might therefore make a very strong case. It is therefore of uttermost importance that these data be stored under a jurisdiction that does not allow data to be seized for court cases or criminalisation of HIV-positive individuals based on their sexual orientation.

Summary

Analysis of sequence data is a powerful tool to study epidemiological and clinical questions if combined with informative metadata. NGS sequencing has opened the door for more sophisticated and precise analyses. With the cost of sequencing continuing to go down, sequence information will become a common source of data in epidemiological and clinical studies.

Glossary

Consensus sequence: Sequence produced from multiple reads of the same sample with the most common base being reported for each position. It is possible that not a single RNA

molecule in the sample is actually identical to the consensus sequence, but it gives a good overall overview of sequences present in the sample

Haplotype: Sequence of a single RNA or DNA molecule present in a sample. Haplotypes are currently different to generate, but will become more frequent with new sequencing methods being developed.

Minority variant: Single nucleotide polymorphism that is not represented in the consensus sequence.

Sequence coverage: Fraction of the genome that was successfully sequenced. Confusingly, “coverage” is sometimes also used interchangeably with sequence depth.

Sequencing depth: Minimum number of times that all parts of a given sequence have been sequenced.